Change-point estimation in the multivariate model taking into account the dependence: Application to the vegetative development of oilseed rape

Vincent Brault, C. Lévy-Leduc, Amélie Mathieu, Alexandra Jullien

Change-point estimation in the multivariate model taking into account the dependence: Application to the vegetative development of oilseed rape

V. Brault[1], C. Lévy-Leduc[2], A. Mathieu[3] and A. Jullien[3]

[1] Univ. Grenoble Alpes, CNRS, LJK,

38000 Grenoble, France

[2] UMR MIA-Paris, INRA, AgroParisTech, Université Paris-Saclay,

75005, Paris, France

[3] UMR ECOSYS, INRA, AgroParisTech, Université Paris-Saclay,

78850, Thiverval-Grignon, France

June 8, 2018

Abstract

In this paper, we address the change-point estimation issue in multivariate observations which consist in functions having piecewise constant first derivatives corrupted by some additional noise. We propose to solve this problem by rewriting it as a variable selection issue in a sparse multivariate linear model. Moreover, the methodology that we propose takes into account the dependence that may exist within the multivariate observations. Then, the performance of our approach is assessed through some numerical experiments and compared

to other alternative and classical methods. Finally, we apply our methodology to experimental data in order to study the vegetative development of oilseed rape. The evolution of the number of leaves of oilseed rape can be modeled as a function having piecewise constant first derivatives corrupted by some additional noise where the change-points correspond to key times in the plant phenology. Our novel estimation method increases the accuracy of the change-point estimation in comparison with classical approaches. Moreover, we show that the parameters of the covariance matrix depend on the level of competition between plants.

KEYWORDS: Multivariate models; change-point estimation; variable selection; dependence; application to oilseed rape; phyllochron

1. INTRODUCTION

Vegetative development of crops conditions the leaf surfaces which insure the captation of light for photosynthesis. At the individual plant scale, the leaf appearance rate is a key factor of this development because it drives the settings of the number of leaves that is highly correlated to the leaf surface at the crop scale (Nanda, Bhargava & Rawson 1995) and *in fine* to the crop yield (Morrison & Mcvetty 1991; Diepenbrock 2000). Recent works have shown that leaf appearance rate may also vary highly with plant-plant competition within the crop (Baey & Cournède 2011) and explains an important part of the leaf surface variability under high pressure of competition far before the variation in the individual leaf surface (Baldissera, Frak, de Faccio Carvalho P.C. & Louarn 2014). For these reasons, the leaf appearance rate is required by many plant models that integrates physiological processes of plant growth (Evers, Vos, Fournier, Andrieu, Chelle & Struik 2005; Jullien, Mathieu, Allirand, Pinet, de Reffye, Cournede & Ney 2011). Studying the variations of these variables of plant development according to growing conditions makes it possible to improve the parameterization of plant models which will in turn increase the accuracy of model simulations and predictions (Gabrielle, Denoroy, Gosse, Justes & Andersen 1998).

In agronomy, the evolution of the number of leaves is usually modelled as a linear function of the thermal time expressed in degree.day, that is the sum of the daily temperature above a base temperature (Bonhomme 2000). Experimental assessment of the relationships is carried out by recording the number of leaves emerged once to twice a week during the whole plant growth and the rate of leaf emergence is computed as the slope of the relationship between the number of leaves and the thermal time. For many species, this relationship is linear and the inverse of the slope is called the phyllochron, that is the time between the appearance of two leaves (Rickman & Klepper 1995). For some species, the number of leaves on the main stem fits with a piecewise linear continuous function with two change-points as observed on wheat (Baker, Allen, Boote, Jones & Jones 1990), rice (Tivet 2000; de Raissac, Audebert, Roques & Bolomier 2004), beetroot (Lemaire, Maupas, Cournede & de Reffye 2008) and rapeseed (Jullien et al. 2011; Gomez & Miralles 1990).

Each change-point is a key time in plant phenology and marks the beginning of a new developmental phase. As far as oilseed rape is concerned, the first phase corresponds to the rosette stage (Miralles, Ferro & Slafer 2001). The change-point had been manually estimated at 610 degree.day at the middle of January in the region of the north of France (Jullien et al. 2011). In the second phase, the phyllochron is reduced (acceleration of the leaf appearance rate) and the second change-point is the anthesis developmental stage (end of leaf appearance and beginning of flower development). The last phase is the reproductive phase where the appearance of flowers replaces those of leaves.

There is a relationship between the timing at which the change in the rate of leaf emergence occured and the final number of leaves (Miralles et al. 2001). The duration of the two phases depends on several factors such as sowing dates or crop density (Miralles et al. 2001; Morrison, Mcvetty & R. 1990). However, the causes and timing of these changes are still not clear-cut. For oilseed rape, (Tittonel 1990) found a link between the date of the change-point and the floral transition on the plant apical meristem while (Miralles et al. 2001) demonstrated that it could not be associated to any particular leaf number. In rice, the second phase starts with the beginning of the stem elongation (de Raissac

et al. 2004) and in beetroot, (Lemaire et al. 2008) highlighted a relationship between the date of the change-point and the competition level. All these reasons justify all the more the necessity of determining exactly the dates of the change-points.

As indicated by (Lemaire et al. 2008; Morrison et al. 1990), the parameters of the dynamics of the number of leaves (rate and change-point) may vary in particular with the plant density. We thus hypothesized that the change-points could be indicators of the competition between plants. We defined an experimental design for winter oilseed rape with different levels of competition generated by different densities and heterogeneity in initial plant size.

In the previously cited studies, change-points were estimated manually or fitted with a linear regression. However, to address this question, we need a method efficient to estimate accurately the change-points and sensitive enough to detect differences between growing conditions. An abundant literature is dedicated to the change-point detection issue for univariate piecewise constant observations corrupted with additive noise both from a theoretical and practical point of view. For a review on the change-point detection field, we refer the reader to (Carlstein, Muller & Siegmund 1994). From a practical point of view, the standard approach for estimating the change-point locations is based on least-square fitting, performed via a dynamic programming algorithm (DP). Indeed, for a given number of change-points K, the dynamic programming algorithm, proposed by (Bellman 1961), takes advantage of the intrinsic additive nature of the least-square objective to recursively compute the optimal change-points locations with a complexity of $O(Kn^2)$ in time, see (Auger & Lawrence 1989) and (Kay 1993). This complexity has recently been improved by (Killick, Fearnhead & Eckley 2012), (Rigaill 2015) and (Maidstone, Hocking, Rigaill & Fearnhead 2016) in some specific cases. A different route to reducing the computational complexity of the multiple change-point detection problem in the univariate case is considered in (Harchaoui & Lévy-Leduc 2010) who consider the least-squares criterion with a total variation penalty, which enables them to use the LARS algorithm of (Efron, Hastie, Johnstone & Tibshirani 2004). Other penalties are proposed in (Ng, Lee & Lee 2018). Another

very popular approach in the one-dimensional case is the Binary Segmentation method proposed by (Scott & Knott 1974) and more recently the Wild Binary Segmentation approach proposed by (Fryzlewicz 2014).

A large literature is also dedicated to change-point detection and estimation in the very general multivariate setting. To list but a few, (Bai 2010) proposed a change-point estimation method in the case where it is assumed that a change in the mean has taken place in each series at an unknown common point. (Horvath & Huskova 2012) proposed a change-point detection approach in the case where the changes occur in the mean and where there is some dependence within each series and not among the different series. (Cho & Fryzlewicz 2015) devised a parametric approach for identifying multiple change-points in the second-order structure of a multivariate (possibly high dimensional) time series based on localized periodograms and cross-periodograms computed on the original multivariate time series.

In this paper, we propose a novel statistical method to estimate the two change-points between the three developmental phases which can be seen as piecewise linear continuous functions corrupted by some additive noise taking into account the dependence that may exist between the different plants. More precisely, we propose to model the number of leaves for the different plants at the different thermal times as a sparse multivariate linear model where the boundaries between the different development stages correspond to the positions of the non null coefficients in the multivariate linear model. Further details on this modeling are given in Section 2.1.

Our contribution then consists in modeling, estimating and removing the dependence (also called "whitening") that may exist between the different plants. The corresponding "whitening" strategy is precisely described in Section 2.2. Then, after a vectorization of the data, the Lasso criterion proposed by (Tibshirani 1996) is applied for finding the positions of the non null coefficients after having applied a preconditioning described in (Jia & Rohe 2015) to the "whitened" observations. Our approach can thus be seen as an extension of the methodology proposed by (Harchaoui & Lévy-Leduc 2010) and (Tibshirani 2014). On the

one hand, none of these approaches can deal with multivariate observations and take into account the dependence that may exist between the different plants. On the other hand, we can deal with the estimation of change-points in observations modeled as piecewise linear continuous functions instead of piecewise constant functions, which is the case considered in (Harchaoui & Lévy-Leduc 2010).

The paper is organized as follows. We first present in Section 2 the statistical modeling and the statistical inference that we propose. Then, in Section 3, we provide some numerical experiments to investigate the statistical performance of our approach which is compared to classical univariate change-point estimation methods. Finally, in Section 4, we apply our methodology to data acquired on winter oilseed rape plants.

2. STATISTICAL FRAMEWORK

2.1 Statistical modeling

Let $\boldsymbol{Y}$ be the $n \times K$ observation matrix such that each column corresponds to the number of leaves for the different dates of observations in thermal time for a given plant. Hence, n corresponds to the total number of observation dates and K to the number of plants. According to (Jullien et al. 2011), the number of leaves can be modeled as a function of the thermal time having piecewise constant first derivatives. Thus, estimating the thermal times corresponding to the boundaries of the different development stages amounts to finding the thermal times at which the slopes of this function changes. In order to estimate the thermal times corresponding to the boundaries of the different development stages, we shall use the following modeling:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}. \tag{1}$$

In (1), $\boldsymbol{X} = (X_{i,j})_{1 \leq i \leq n, 1 \leq j \leq n}$ such that

$$\boldsymbol{X} = \begin{pmatrix} t_1 & 0 & & \dots & 0 \\ t_2 & t_2 - t_1 & 0 & \dots & 0 \\ \vdots & & \ddots & & 0 \\ t_n & t_n - t_1 & 0 & \dots & t_n - t_{n-1} \end{pmatrix},$$

where the t_k's correspond to the different thermal times recorded and $\boldsymbol{E}$ is the $n \times K$ random error matrix. In (1), $\boldsymbol{B}$ is a sparse matrix such that the positions of its non-null coefficients correspond to the change-point positions. Thanks to this modeling, finding the positions of the non-null coefficients in $\boldsymbol{B}$ allows us to find the thermal times at which the changes occur in the slope of the function having piecewise constant first derivatives. Note that, as in (Harchaoui & Lévy-Leduc 2010), we modeled the change-point detection problem as an estimation issue in a sparse linear model. However, there are some differences. First, we extended their modeling to the multivariate case since, here, $\boldsymbol{Y}$ and $\boldsymbol{E}$ are matrices and not vectors. Second, the matrix $\boldsymbol{X}$ also changed: it is not anymore a lower triangular matrix with nonzero elements equal to one since this design matrix is dedicated to the detection of changes in piecewise constant observations.

In this paper, we shall pay a special attention to the dependence that may exist between the different columns of $\boldsymbol{E}$, namely between the different plants. More precisely, we shall assume that the rows of $\boldsymbol{E}$ are independent, identically distributed and such that

$$(E_{i,1}, E_{i,2}, \ldots, E_{i,q}) \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_K).$$

In the following, the covariance matrix $\boldsymbol{\Sigma}_K$ will be chosen in order to take into account the spatial dependence that may exist among the different plants. Hence, $\boldsymbol{\Sigma}_K = (\Sigma_{i,j})_{1 \leq i,j \leq K}$ will be defined by

$$\Sigma_{i,j} = \sigma^2 \exp\left(-\frac{\|P_i - P_j\|^2}{2\ell^2}\right), \tag{2}$$

where σ is a real parameter, $\|P_i - P_j\|$ denotes the euclidean distance in $\mathbb{R}^2$ between the plants i and j computed according to their positions in the containers and ℓ is in $(0, 1)$.

Our goal will be to devise a methodology for estimating the two change-points corresponding to the boundaries of the different stages of development in the number of leaves as a function of the thermal time for each plant of a given container using the modeling described in (1) and (2).

2.2 Statistical inference

The methodology that we propose is a Lasso based approach thus we first briefly recall the usual framework in which the Lasso approach is used.

Let us consider a high-dimensional linear model of the following form

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}, \tag{3}$$

where $\mathcal{Y}$, $\mathcal{B}$ and $\mathcal{E}$ are vectors. Note that, in high-dimensional linear models, the matrix $\mathcal{X}$ has usually more columns than rows which means that the number of variables is larger than the number of observations but $\mathcal{B}$ is usually a sparse vector, namely it contains a lot of null components. In such models a very popular approach initially proposed by (Tibshirani 1996) consists in using the Least Absolute Shrinkage eStimatOr (LASSO) criterion for estimating $\mathcal{B}$ defined as follows for a positive λ:

$$\widehat{\mathcal{B}}(\lambda) = \text{Argmin}_{\mathcal{B}} \left\{ \|\mathcal{Y} - \mathcal{X}\mathcal{B}\|_2^2 + \lambda \|\mathcal{B}\|_1 \right\}, \tag{4}$$

where, for $u = (u_1, \ldots, u_n)$, $\|u\|_2^2 = \sum_{i=1}^n u_i^2$ and $\|u\|_1 = \sum_{i=1}^n |u_i|$, which is usually called the ℓ_1-norm of the vector u. Observe that the first term of (4) is the classical least-squares criterion and that $\lambda\|\mathcal{B}\|_1$ can be seen as a penalty term. The interest of such a criterion is the sparsity enforcing property of the ℓ_1-norm ensuring that the number of non-zero components of the estimator $\widehat{\mathcal{B}}$ of $\mathcal{B}$ is small for large enough values of λ. Such a criterion is very relevant in our framework since the problem of finding the change-points boils down to finding the non null coefficients in the matrix $\boldsymbol{B}$.

This methodology cannot be directly applied to our model since we have to deal with matrices and not with vectors. However, by using the following trick, Model (1) can be rewritten as in (3) where $\mathcal{Y}$, $\mathcal{B}$ and $\mathcal{E}$ are vectors of size nK.

Let $vec(\boldsymbol{A})$ denote the vectorization of the matrix $\boldsymbol{A}$ formed by stacking the columns of $\boldsymbol{A}$ into a single column vector. Then, with

$$\mathcal{Y} = vec(\boldsymbol{Y}),\ \mathcal{B} = vec(\boldsymbol{B}) \text{ and } \mathcal{E} = vec(\boldsymbol{E}), \tag{5}$$

we get (3) with

$$\mathcal{X} = \mathrm{I}_q \otimes \boldsymbol{X}, \tag{6}$$

where $\otimes$ denotes the Kronecker product.

Let us now summarize the methodology that we propose for estimating the changes in the slopes:

- First step: Estimation of the random error matrix $\boldsymbol{E}$ by $\widehat{\boldsymbol{E}}$ using a Lasso based approach.
- Second step: Estimation of $\boldsymbol{\Sigma}_K$ by $\widehat{\boldsymbol{\Sigma}}_K$ thanks to an estimation of σ and ℓ in (2).
- Third step: Thanks to the estimator $\widehat{\boldsymbol{\Sigma}}_K$, transforming the data in order to remove the dependence between the columns of $\boldsymbol{Y}$. Such a transformation will be called "whitening" hereafter.
- Fourth step: Estimation of $\boldsymbol{B}$ using a Lasso based approach.

These different steps are described hereafter.

Estimation of the error matrix $\boldsymbol{E}$ and of $\boldsymbol{\Sigma}_K$ (first and second steps). In order to obtain an estimation of $\widehat{\boldsymbol{E}}$, we use (4) with $\mathcal{Y}$ and $\mathcal{X}$ defined in (5) and (6), respectively and where λ is chosen by cross-validation (CV). We thus obtain $\widehat{\mathcal{E}}$ as follows:

$$\widehat{\mathcal{E}} = \mathcal{Y} - \mathcal{X}\widehat{\mathcal{B}}(\lambda_{\mathrm{CV}}),$$

and hence $\widehat{\boldsymbol{E}}$ by using that $\widehat{\mathcal{E}} = vec(\widehat{\boldsymbol{E}})$. The estimations of ℓ and σ in (2) are then obtained thanks to the maximum likelihood approach.

Note that this approach provides better results than the one consisting in estimating the change-points within each column of $\boldsymbol{Y}$ by using a maximum likelihood approach in order to have an estimation of $\boldsymbol{E}$ and then to estimate ℓ and σ by a maximum likelihood approach.

Whitening step (third step). In order to remove the dependence that may exist between the columns of $\boldsymbol{Y}$, we shall use the following transformation:

$$\boldsymbol{Y}\ \boldsymbol{\Sigma}_K^{-1/2} = \boldsymbol{X}\boldsymbol{B}\ \boldsymbol{\Sigma}_K^{-1/2} + \boldsymbol{E}\ \boldsymbol{\Sigma}_K^{-1/2}. \tag{7}$$

Since $\boldsymbol{\Sigma}_K$ is in general unknown, we shall replace it by its estimator $\widehat{\boldsymbol{\Sigma}}_K$ defined by (2) where ℓ and σ are replaced by their estimator obtained in the second step of our methodology. We shall thus consider the following transformation:

$$\boldsymbol{Y}\ \widehat{\boldsymbol{\Sigma}}_K^{-1/2} = \boldsymbol{X}\boldsymbol{B}\ \widehat{\boldsymbol{\Sigma}}_K^{-1/2} + \boldsymbol{E}\ \widehat{\boldsymbol{\Sigma}}_K^{-1/2}. \tag{8}$$

Estimation of $\boldsymbol{B}$ (fourth step). In order to take into account the dependence between the columns of $\boldsymbol{Y}$ we propose using a modified version of the standard Lasso criterion (4). More precisely, by applying the *vec* operator to (8), we get (3) where

$$\mathcal{Y} = vec(\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}_K^{-1/2}),\ \mathcal{X} = (\widehat{\boldsymbol{\Sigma}}_K^{-1/2})' \otimes \boldsymbol{X},\ \mathcal{B} = vec(\boldsymbol{B}) \text{ and } \mathcal{E} = vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}_K^{-1/2}). \tag{9}$$

Hence, retrieving the positions of the non null components in $\widehat{\mathcal{B}}$ defined in (4) with $\mathcal{Y}$ and $\mathcal{X}$ previously defined provides the estimators of the change-point locations.

Following (Jia & Rohe 2015) preconditioning $\mathcal{X}$ by using a Puffer transformation may improve the ability of the Lasso criterion to properly retrieve the null and non-null positions in $\boldsymbol{B}$. Hence, we shall use a transformation on Model (3) where $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{B}$ and $\mathcal{E}$ are defined in (9). Let

$$\mathcal{X} = UDV'$$

denote the SVD of $\mathcal{X}$ where U and V are orthogonal matrices and D is a diagonal matrix containing the singular values of $\mathcal{X}$. Let also

$$F = UD^{-1}U'$$

denote the Puffer transformation defined in Section 2.1 of (Jia & Rohe 2015). Then, instead of applying the Lasso criterion directly on (3) with $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{B}$ and $\mathcal{E}$ defined in (9), we shall apply it on:

$$F\mathcal{Y} = F\mathcal{X}\mathcal{B} + F\mathcal{E}. \tag{10}$$

With such a transformation, $F\mathcal{E}$ is a centered Gaussian random vector having a covariance matrix equal to $UD^{-2}U'$, the diagonal values of D^{-2} may thus be very large for very small singular values of D. For this reason, we propose in the following to keep only in D the m largest singular values of $\mathcal{X}$, hence, instead of (10), we shall consider

$$F_m\mathcal{Y} = F_m\mathcal{X}\mathcal{B} + F_m\mathcal{E}, \tag{11}$$

where

$$F_m = UD_m^{-1}U',$$

where D_m^{-1} is a diagonal matrix having on its diagonal the inverse of the m largest singular values of $\mathcal{X}$ and 0 on the other entries of the diagonal. We shall explain in Section 3 how to choose m in practical situations.

The parameter λ in the Lasso criterion (4) is chosen by cross-validation. Then, the estimated positions of the two change-points for each plant k correspond either to the two largest values of $|\widehat{\boldsymbol{B}}_{\cdot,k}|$ (approach called "two max" in Section 3) or to the largest and smallest values of $\widehat{\boldsymbol{B}}_{\cdot,k}$ (approach called "min/max" in Section 3), where $\widehat{\boldsymbol{B}}_{\cdot,k}$ denotes the kth column of $\widehat{\boldsymbol{B}}$. We refer the reader to Section 3 for a further comparison of these estimation approaches.

3. NUMERICAL EXPERIMENTS

To assess the performance of our methodology, we generated observations $\boldsymbol{Y}$ according to Model (1) with $n = 28$, $K = 42$ and Σ_K defined in (2) for ℓ in $\{0.2, 0.5, 0.8\}$ and σ in $\{1, 2, 5, 10\}$. In the following, we shall denote by $t_1^{\star,k}$ and $t_2^{\star,k}$ the positions of the change-points for the plant k.

Note that we have chosen the values of the parameters n and K in order to be as close as possible to the real data that we plan to analyze in Section 4.

The performance of our approach described in Steps 1 and 2 for estimating ℓ and σ defined in (2) are displayed in Figure 1 given in the Supplementary material. We can see from this figure that the best estimations of ℓ are obtained for $\ell = 0.5$ and for the other

values of ℓ the best estimations of ℓ are obtained for large values of σ. We observe from the right part of Figure 1 that σ is generally underestimated except for small values of σ ($\sigma = 1$). However, we shall see in the following that even if ℓ and σ are not very precisely estimated the estimation of the $t_1^{\star,k}$'s and $t_2^{\star,k}$'s is not altered.

In Figures 2, 3 and 4 given in the Supplementary material, our procedure is compared with other methodologies for estimating the $t_1^{\star,k}$'s. These figures display for different values of σ and ℓ the boxplots of the empirical mean over k of $|\widehat{t_1}^k - t_1^{\star,k}|$ as a function of m for 100 replications obtained with procedures containing a whitening step such as ours and with procedures which do not contain such a stage. More precisely, the tested procedures are:

- (ML) the maximum likelihood approach which is designed for finding two change-points in the slope of a function having a piecewise constant first derivative without taking into account the dependence that exists between the columns of $\boldsymbol{Y}$.
- (M1) Lasso criterion applied to (11) with $\widehat{\Sigma}_K^{-1/2} = \mathrm{Id}_{\mathbb{R}^K}$
- (M2) Lasso criterion applied to (11) with $\widehat{\Sigma}_K^{-1/2} = \Sigma_K^{-1/2}$, which never occurs in practice
- (M3) Lasso criterion applied to (11) with $\widehat{\Sigma}_K^{-1/2}$ defined by (2) where ℓ and σ are replaced by their estimators obtained in the first and second step of our method (our approach).

For each of the last three approaches, $\widehat{t_1}^k$ or $\widehat{t_2}^k$ correspond either to the two largest values of $|\widehat{\boldsymbol{B}}_{\cdot,k}|$ (two max) or to the largest and smallest values of $\widehat{\boldsymbol{B}}_{\cdot,k}$ (min/max), where $\widehat{\boldsymbol{B}}_{\cdot,k}$ denotes the kth column of $\widehat{\boldsymbol{B}}$.

Figures 5, 6 and 7 of the Supplementary material display similar results for the $t_2^{\star,k}$'s.

We can see from these figures that the best results are obtained by our approach (M3) with the (min/max) method for well chosen values of m and that they are very close to the method using the knowledge of Σ_K which is assumed to be unknown in our strategy.

Let us now focus on the choice of m. After some investigation, we observe that finding the best value of m is not an easy task since it is very unstable. Thus, instead of selecting

the m largest singular values of $\mathcal{X}$, we propose to keep the singular values larger than a given threshold. In the following, we provide the performance of (M3) when the threshold is equal to 2 for the two max and min/max approaches and compare them with (ML). More precisely, Figure 8 given in the Supplementary material displays a comparison of (ML) to (M3) with two max and min/max. We observe from this figure that the only regions in which (ML) provides better results than (M3) with two max are those where σ is small. For (M3) with min/max we do not observe such a phenomenon. Our approach indeed exhibits particularly good performance for estimating the $t_2^{\star,k}$'s and performance equivalent to (ML) for small values of ℓ. Further comparisons are given in Table 1 which gives the percentage of times where the procedures (M3) two max and (M3) min/max performs better than (ML).

Note that we only compared our approaches with (ML) since it provides better results than the generalized least-squares approach. Indeed, Figure 1 displays the boxplots of $|\widehat{t_1}^k - t_1^{\star,k}|$ and $|\widehat{t_2}^k - t_2^{\star,k}|$ obtained with (ML) and with the maximum likelihood approach applied to the whitened observations $\boldsymbol{Y}\boldsymbol{\Sigma}_K^{-1/2}$, denoted by (GLS) in the following, for different values of ℓ and σ defined in (2). We can see from Figure 1 that the performance of (GLS) are on a par with those of (ML) for small values of ℓ and that they are altered for large values of ℓ. This comes from the fact that for removing the spatial dependence within the plants (dependence within the columns of $\boldsymbol{Y}$), $\boldsymbol{Y}$ has to be multiplied on the right by $\boldsymbol{\Sigma}_K^{-1/2}$, see Equation (7), which changes the values of $\boldsymbol{B}$, contrary to our approach, where only the design matrix is modified.

4. APPLICATION TO THE VEGETATIVE DEVELOPMENT OF OILSEED RAPE

In this section, we apply the methodology devised in Section 3 to the data acquired on winter oilseed rape plants under agronomic conditions.

4.1 Description of field experiments

Field experiments were conducted at the experimental unit of INRA Thiverval-Grignon (France, N 48 ° 51′20″ E 1 ° 56′25”) on the winter oilseed rape cultivar Pollen. Seeds were sown in individual pots on August 31st and plants were transplanted about two weeks later

			Two max				min/max			
			σ				σ			
			1	2	5	10	1	2	5	10
t_1	ℓ	0.2	1%	31%	91%	99%	41%	36%	42%	51%
		0.5	4%	24%	86%	96%	56%	29%	56%	57%
		0.8	2%	25%	88%	98%	75%	80%	94%	100%
t_2	ℓ	0.2	8%	48%	98%	100%	56%	83%	99%	100%
		0.5	5%	56%	99%	100%	53%	86%	100%	100%
		0.8	7%	64%	99%	100%	83%	95%	100%	100%

Table 1: Percentage of times where the procedures (M3) two max (left) and (M3) min/max (right) are better than (ML) for estimating the $t_1^{\star,k}$'s (first row) and the $t_2^{\star,k}$'s (second row) for ℓ in $\{0.2, 0.5, 0.8\}$ and σ in $\{1, 2, 5, 10\}$.

into eight containers of 1.2 meter square. We hypothethised that density and heterogeneity in initial plant size may modify the competition between plants and the individual dynamics in number of leaves. Therefore three different treatments were carried out, each applied to one to three plant container: high plant density homogeneous (HO), high plant density heterogeneous (HE) and low density homogeneous (LD), see Table 1 of the Supplementary

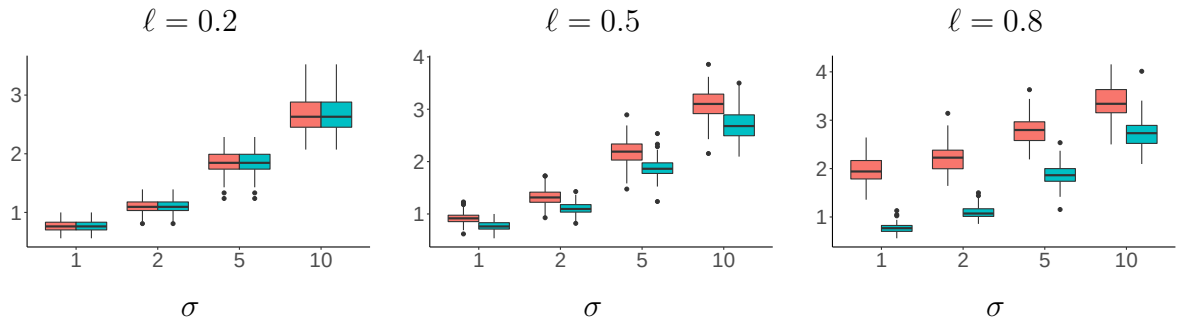

Figure 1: Boxplots of $|\widehat{t}_1^{\,k} - t_1^{\star,k}|$ and $|\widehat{t}_2^{\,k} - t_2^{\star,k}|$ for the different values of k for the (ML) (blue) and (GLS) (red) approaches for different values of σ and ℓ.

material. The highest density was 35 plants.m^2 and the lowest one was 20 plants.m^2. For the homogeneous treatment plants were selected at transplantation in order to be as similar as possible (2 leaves, similar leaf surface) while for the heterogeneous treatment plants were seperated into three categories of plant leaf surface (small, middle, big) and mixed in equivalent proportion. For each of the three treatments, containers were harvested at two dates : 780 °D (Early) and 923 °D (Late).

Thermal time was computed with the daily average temperature of the meteorological station of Thiverval-Grignon, and base temperature was set to 4.5 °C, which is a classical value for winter oilseed rape (Gabrielle et al. 1998).

4.2 Results on real data

In order to check the presence of dependence between the different plants, the boxplots of the p-values of the Pearson correlation tests for the different containers before and after the whitening step are displayed in Figure 2. We can see from this figure that there exists some dependence between the plants and that our whitening approach almost removed this dependence.

Table 2 displays the estimations of ℓ and σ defined in (2) for the eight plant containers. The estimations of σ are smaller than 1 and the estimations of ℓ are close to 0.7. We observe that σ is slightly higher for Containers $1, 5$ and 8 *i.e.* high density and late harvest date,

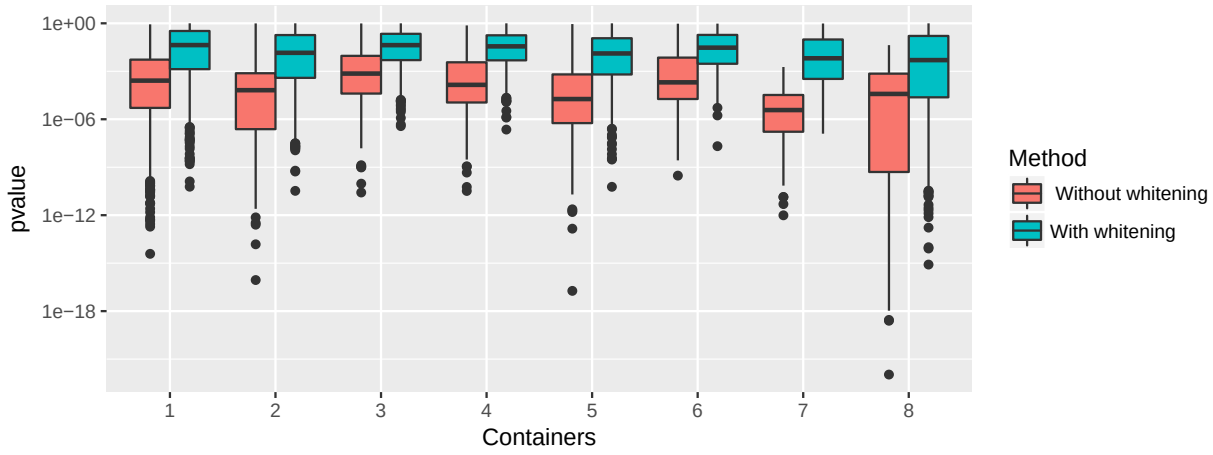

Figure 2: Boxplots of the p-values of the Pearson correlation tests using a logarithmic scale for the different containers before and after the whitening step.

corresponding to a higher level of competition between plants. However, σ is only 0.38 for Container 2 that followed the same conditions. Within the HO_Late treatment, plants of Container 1 had in average higher initial leaf surface areas than plants of Containers 2 and 5 (data not shown) which may explain a higher value of σ. Indeed, the higher the leaf area, the greater the competition for light.

Based on Table 1, our approach (M3) with min/max provides better results than (M3) with two max in the range of values of σ and ℓ. Moreover, the performance of our approach (M3) with min/max are at least comparable with those of (ML) or even better. Figure 3 displays the boxplots of the $\widehat{t}_1^{\,k}$'s and $\widehat{t}_2^{\,k}$'s obtained by the (ML) and (M3) approaches for the different containers. We can see from this figure that the estimations provided by (ML) have a very high variability contrary to our approach (M3). The latter is therefore preferable to have a more precise estimation of the change-points. Taking into account the spatial correlation improves the estimation of the parameters.

The (M3) method with two max returns more variable results than the (M3) with min/max. However, the estimations given by the (M3) method with two max are more coherent from an agronomical point of view. Indeed, the first estimated change-point with the (M3) two max was around 600 to 620 $^\circ$D with variation according to the different

treatments which is in agreement with the biological model. The value is close to the one found by (Jullien et al. 2011) (610 °D) and coherent with the datation of floral initiation in (Miralles et al. 2001). The (M3) method with two max provides later estimations of $\widehat{t_1}^k$'s, and very close to $\widehat{t_2}^k$'s, which contradicts the agronomical knowledge. As far as the second change-point is concerned, the two (M3) methods gave similar estimations. This can be explained by the fact that the min/max method looks for a strongly positive slope difference followed by a strongly negative slope difference whereas the two max method looks for two high slope differences in absolute value. In our case study, the slopes are very high in the second phase compared to the first phase, and the min/max method tends to locate the two change-points quite close during this second phase. This is illustrated by Figure 4 that compares the results of the three methods on the data of individual plants. Each subplot shows the number of leaves as a function of the thermal time for a given plant and the $\widehat{t_1}^k$'s and $\widehat{t_2}^k$'s for the methods (ML), (M3) with two max and (M3) with min/max. Individual plant data are available for the other containers upon request.

To conclude on methods comparison, the (M3) with two max is both more precise than (ML) and more coherent with the biological model than the (M3) with min/max, at least for winter oilseed rape.

For species with different patterns of variations in phyllochron between the two phases, the method (M3) with min/max could give better results. For instance, for WOSR the slope increases between the two phases whereas it decreases for wheat (Miralles et al. 2001) or beetroot (Lemaire et al. 2008).

The (M3) methods with two max provides similar values for the two change-points for the three containers of treatment HO_Late: between 600 °D and 650 °D for the first change-point and between 700 °D and 750 °D for the second one. This shows that despite different values of σ (see Table 2) revealing different levels of biological variability within a container, the estimation of the change points were stable.

While this method is robust to the biological variability, it is very sensitive to the last number of measurement dates. In Figure 3, we observe that treatments harvested earlier

have a change-point around 580 °D (Containers 3, 4 and 6) which is lower than the average value for the other treatments (about 630 °D). To a lesser extent, this is also the case for the second change-point. From a methodological point of view, it confirms that it is crucial to have the full series of measurements to estimate precisely the change point. This default could be improved by increasing the frequency of measurements, particularly during the second phase (600 to 800 °D). The three containers present also a smaller inter-plants variations which can be due to a smaller number of observations (hence a smaller number of possible positions for the change-point).

Heterogenous size at transplantation and density appeared to have little effect on the estimation of the change-point. Estimations of change-points were not modified even if the σ has been shown to be higher for the treatments with a high competition level. This shows that the method was robust to the plant heterogeneity. Or it can also be explained by the fact that the difference in competition levels induced by plant densities and heterogeneity applied were not very contrasted. The method should be further tested on more contrasted situations before to conclude. As an example, (Zhu, Vos, van der Wer, van der Putten & Evers 2014) estimated the phyllochron of maize in crops of different complexity: monoculture or mixed stands associated with wheat with different intercrop. Phyllochron differed according to the crop complexity and it also seemed to affect the change-points according to their data.

Plant container	1	2	3	4	5	6	7	8
$\widehat{\ell}$	0.7	0.64	0.6	0.63	0.66	0.64	0.69	0.63
$\widehat{\sigma}$	0.70	0.38	0.36	0.33	0.54	0.39	0.32	0.71

Table 2: Estimation of ℓ and σ for the different plant containers.

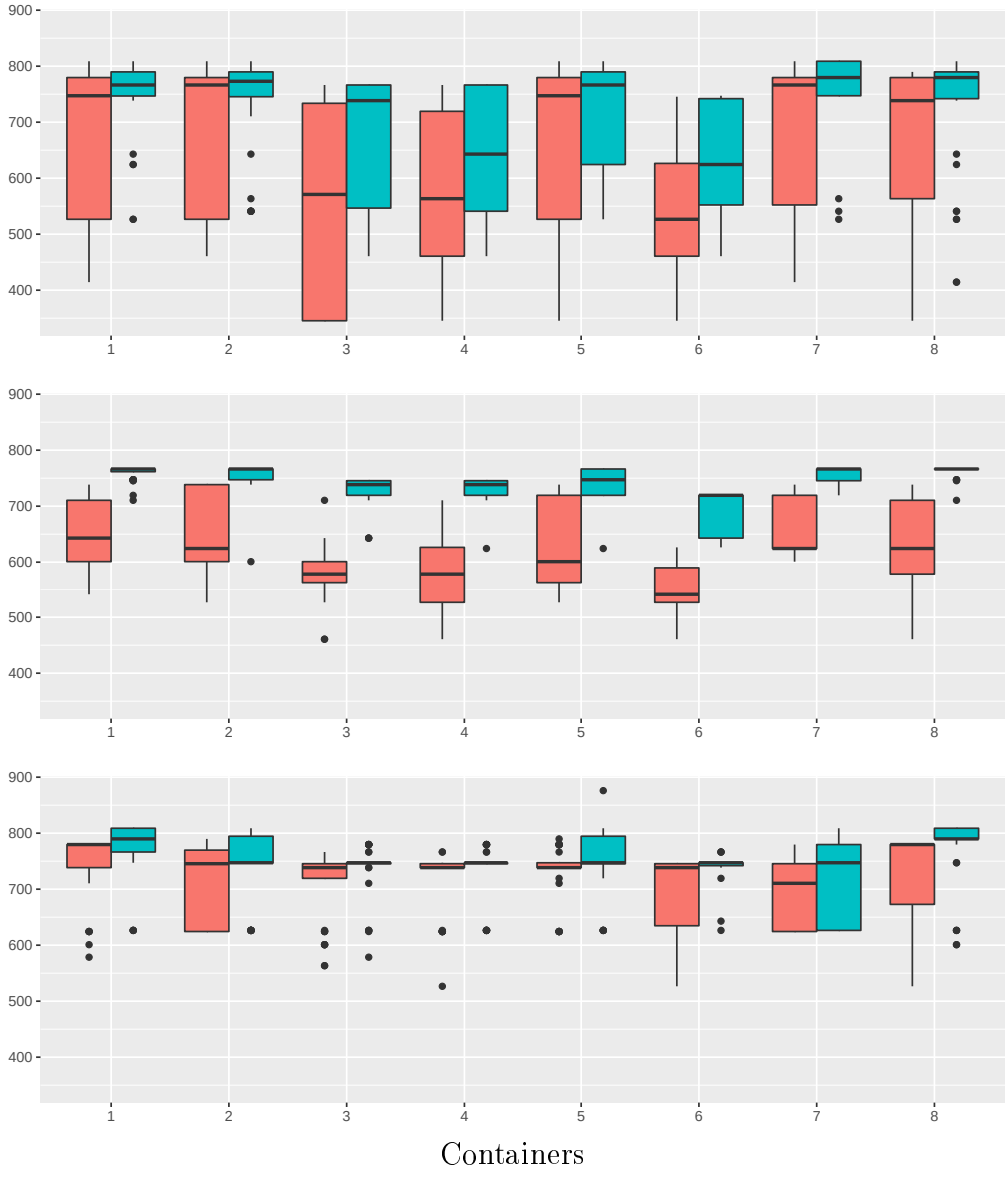

Figure 3: Boxplots of the $\widehat{t_1}^k$'s (red) and $\widehat{t_2}^k$'s (blue) as a function of the plant container (x-axis) for each method: (ML) (top), (M3) with two max (middle) and (M3) with min/max (bottom). Containers 1,2 and 5 are HO_Late treatment; Container 3 is HE_Early treatment, Container 4 is HE_Early, Container 6 is LD_Early, 8 is HE_Late and Container 7 is LD_Late.

5. CONCLUSION

The novel statistical method proposed in this article improves the estimation of the change-points in leaf development models in comparison with the classical methods used in agron-

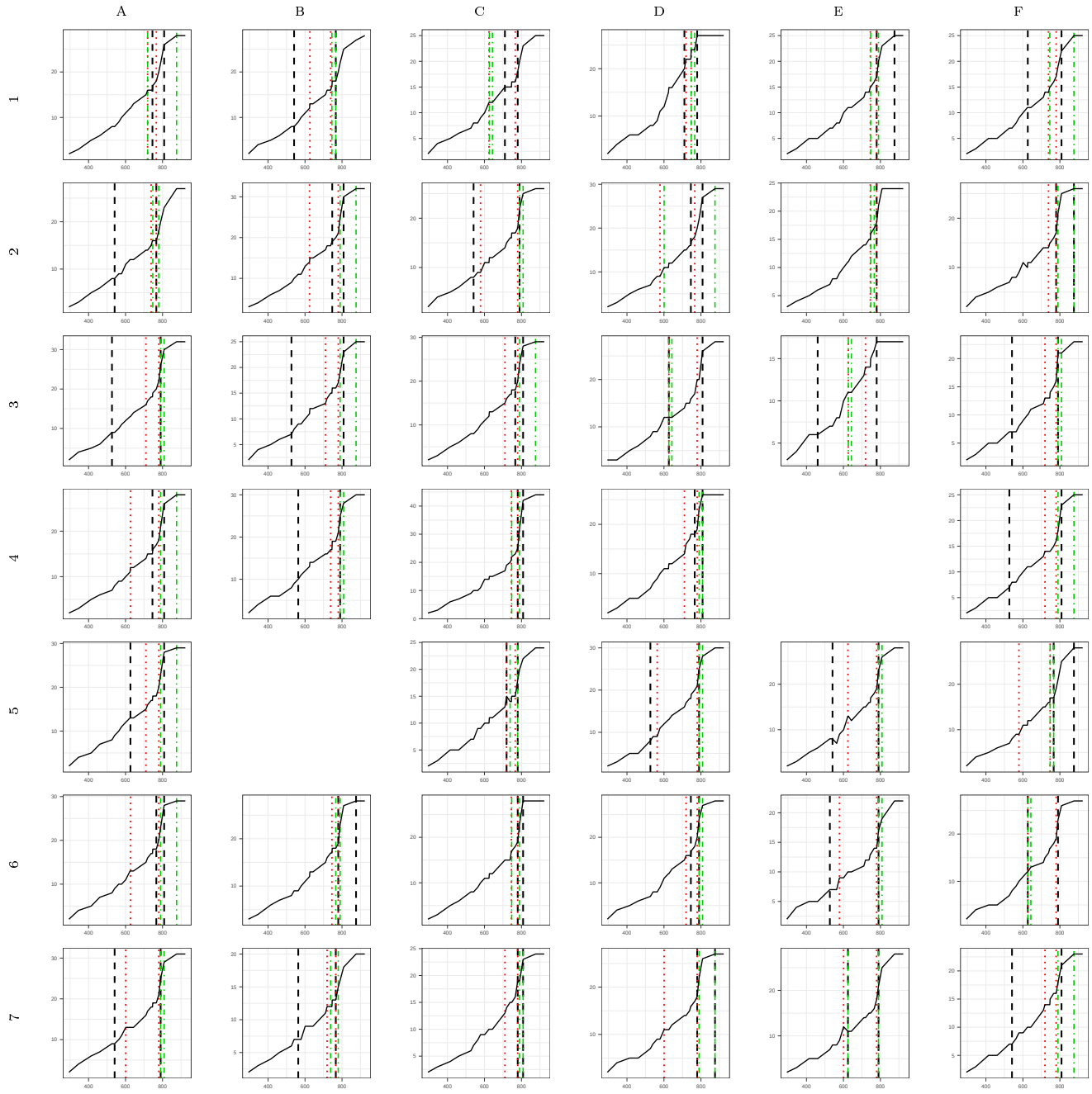

Figure 4: Estimation of the $\widehat{t}_1^{\,k}$'s and the $\widehat{t}_2^{\,k}$'s with $k \in \{A1, \ldots, A7, B1, \ldots, F7\} \backslash \{B5, E4\}$ for Container 1: (M1) (black), (M3) with two max (red) and (M3) with min/max (green). Each plot displays the number of leaves (y-axis) as a function of the thermal time (x-axis) for each plant k, where $k \in \{A1, \ldots, A7, B1, \ldots, F7\} \backslash \{B5, E4\}$.

omy. Covariance matrices estimated on experimental data presented an increasing variability (σ) with increasing levels of competition induced by density or heterogeneity in plant sizes. However, the selected statistical method was robust to this variability and estimations of change-points were similar between treatments around 620 °D for the first one and 750 °D for the second one. The method was sensitive to the time of final plant harvest and thus to the number of measurement dates that should be a particular point of vigilance for the further experiments. In this paper, we used a Lasso criterion where the least-squares part is equivalent to the negative log-likelihood only in the Gaussian case. Since the number of leaves are non negative discrete valued observations it could be interesting to extend our methodology to deal with this case using, for instance, a Poisson distribution. This will be the subject of a future work.

REFERENCES

Auger, I. E., & Lawrence, C. E. (1989), "Algorithms for the optimal identification of segment neighborhoods," Bulletin of Mathematical Biology, 51(1), 39–54.

Baey, C., & Cournède, P. (2011), Using a hierarchical segmented model to assess the dynamics of leaf appearance in plant populations„ in 14th Applied Stochastic Models and Data Analysis International Conference (ASMDA 2011).

Bai, J. (2010), "Common breaks in means and variances for panel data," Journal of Econometrics, 157(1), 78 – 92. Nonlinear and Nonparametric Methods in Econometrics.

Baker, J., Allen, L., Boote, K., Jones, P., & Jones, J. (1990), "Developmental responses of rice to photoperiod and carbon dioxide concentration," Agricultural and Forest Meteorology, 50, 201–210.

Baldissera, T., Frak, E., de Faccio Carvalho P.C., & Louarn, G. (2014), "Plant development controls leaf area expansion in alfalfa plants competing for light," Annals of Botany, 113(1), 145–157.

Bellman, R. (1961), "On the Approximation of Curves by Line Segments Using Dynamic Programming," Commun. ACM, 4(6), 284–286.

Bonhomme, R. (2000), "Bases and limited to using 'degree.day' units," European Journal of Agronomy, 13(1), 1–10.

Carlstein, E., Muller, H. G., & Siegmund, D. (1994), Change-point problems, Hayward: Institute of Mathematical Statistics Lecture Notes.

Cho, H., & Fryzlewicz, P. (2015), "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(2), 475–507.

de Raissac, M., Audebert, A., Roques, S., & Bolomier, J. (2004), Competition between plants affects phenology in rice cultivars,, in New directions for a diverse planet : Proceedings for the 4th International Crop Science Congress, eds. N. Turner, J. Angus, L. Mc Intyre, M. Robertson, A. Borrell, & D. Lloyd, Gosford : Regional Institute.

Diepenbrock, W. (2000), "Yield analysis of winter oilseed rape (*Brassica napus* L.) : a review," Field Crops Research, 67, 35–49.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004), "Least angle regression," The Annals of statistics, 32(2), 407–499.

Evers, J., Vos, J., Fournier, C., Andrieu, B., Chelle, M., & Struik, P. (2005), "Towards a generic architectural model of tillering in Graminae, as exemplified by spring wheat (*Triticum aestivum*)," New Phytologist, 166(3), 801–812.

Fryzlewicz, P. (2014), "Wild binary segmentation for multiple change-point detection," Ann. Statist., 42(6), 2243–2281.

Gabrielle, B., Denoroy, P., Gosse, G., Justes, E., & Andersen, M. (1998), "A model of leaf area development and senescence for winter oilseed rape," Field Crops Research, 57, 209–222.

Gomez, N., & Miralles, D. (1990), "Factors that modify early and late reproductive phases in oilseed rape (Brassica napus L .): Its impact on seed yield and oil content.," Industrial Crops and Products, 34, 1277–1285.

Harchaoui, Z., & Lévy-Leduc, C. (2010), "Multiple Change-Point Estimation With a Total Variation Penalty," Journal of the American Statistical Association, 105(492), 1480–1493.

Horvath, L., & Huskova, M. (2012), "Change-point detection in panel data," Journal of Time Series Analysis, 33(4), 631–648.

Jia, J., & Rohe, K. (2015), "Preconditioning the Lasso for sign consistency," Electron. J. Statist., 9(1), 1150–1172.

Jullien, A., Mathieu, A., Allirand, J., Pinet, A., de Reffye, P., Cournede, P., & Ney, B. (2011), "Characterization of the interactions between architecture and source-sink relationships in winter oilseed rape (Brassica napus) using the GreenLab model," Annals of Botany, 107(5), 765–779.

Kay, S. (1993), Fundamentals of statistical signal processing: detection theory, : Prentice-Hall, Inc.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012), "Optimal Detection of Changepoints With a Linear Computational Cost," Journal of the American Statistical Association, 107(500), 1590–1598.

Lemaire, S., Maupas, F., Cournede, P., & de Reffye, P. (2008), "A morphogenetic crop model for sugar-beet (Beta vulgaris L.)," International Symposium on Crop Modeling and Decision Support: ISCMDS, 5, 19–22.

Maidstone, R., Hocking, T., Rigaill, G., & Fearnhead, P. (2016), "On optimal multiple changepoint algorithms for large data," Statistics and Computing, pp. 1–15.

Miralles, D., Ferro, B., & Slafer, G. (2001), "Developmental responses to sowing date in wheat, barley and rapeseed," Field Crop Research, 71, 211–223.

Morrison, M., & Mcvetty, P. (1991), "Leaf appearance rate of summer rape," Can. J. Plant Sci., 71, 405–412.

Morrison, M., Mcvetty, P., & R., S. (1990), "Effect of altering plant density on growth characteristics of summer rape," Can. J. Plant Sci., 70, 139–149.

Nanda, R., Bhargava, S., & Rawson, H. M. (1995), "Effect of sowing date on rates of leaf appearance , final leaf numbers and areas in Brassica campestris , B . juncea , B . napus and B . carinata," Field Crops Research, 42, 125–134.

Ng, C. T., Lee, W., & Lee, Y. (2018), "Change-point estimators with true identification property," Bernoulli, 24(1), 616–660.

Rickman, R., & Klepper, B. (1995), "The Phyllochron: where do we go in the future?," Crop Science, 35, 44–49.

Rigaill, G. (2015), "A pruned dynamic programming algorithm to recover the best segmentations with 1 to Kmax change-points," Journal de la Société Française de Statistique, 156(4), 180–205.

Scott, A. J., & Knott, M. (1974), "A cluster analysis method for grouping means in the analysis of variance," Biometrics, 30(3), 507–512.

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso," J. Royal. Statist. Soc B., 58(1), 267–288.

Tibshirani, R. J. (2014), "Adaptive piecewise polynomial estimation via trend filtering," Ann. Statist., 42(1), 285–323.

Tittonel, E. (1990), Evènements liés à l'évolution florale chez le colza Brassica napus L. var Oleifera Metzg, PhD thesis, Université Paris Sud, Centre d'Orsay, Paris.

Tivet, F. (2000), Etude des facteurs génotypiques et environnementaux déterminant la mise en place de la surface foliaire chez le riz. Incidence particuliere d'un deficit hydrique, PhD thesis, INA P-G.

Zhu, J., Vos, J., van der Wer, W., van der Putten, P., & Evers, J. (2014), "Early competition shapes maize whole-plant development in mixed stands," Journal of Experimental Botany, 65(2), 641–653.